GÖDEL AGENT: A SELF-REFERENTIAL FRAMEWORK FOR AGENTS RECURSIVELY SELF-IMPROVEMENT

Xunjian Yin, Xinyi Wang, Liangming Pan, Xiaojun Wan, William Yang Wang

Published: October 6, 2024

Problem Statement

How can we create AI agents capable of recursively improving themselves without relying on predefined routines or fixed optimization algorithms?

Methodology

Implementation

1. Use "monkey patching" for dynamic code modification
2. Employ runtime memory techniques for initialization

Self-Improvement Mechanism

1. Recursive self-improvement through a self-referential learning algorithm
2. Use LLMs for autonomous behavior modification

Gödel Machine Inspiration

Based on Gödel machine concept for global optimal solutions
Enables recursive updates and self-reference capabilities

Results and Findings

Performance Comparison

Task: DROP (Reading Comprehension)
Gödel Agent: 80.9% F1 Score
Meta Agent Search: 79.4% F1 Score
Best Hand-Designed: 65.8% F1 Score

Task: MGSM (Mathematics)
Gödel Agent: 64.2% Accuracy
Meta Agent Search: 53.4% Accuracy
Best Hand-Designed: 39.0% Accuracy

Task: GPQA (Graduate-level Science)
Gödel Agent: 34.9% Accuracy
Meta Agent Search: 34.6% Accuracy
Best Hand-Designed: 31.6% Accuracy

Key Takeaways

1. Gödel Agent outperforms manually designed and meta-learning optimized agents
2. Framework demonstrates superior adaptability and efficiency across various tasks
3. Self-referential approach allows for exploration of full agent design space
4. LLM-driven decision making enables creative problem-solving strategies
5. Recursive self-improvement leads to continuous performance gains

Limitations and Future Work

Limitations

1. High complexity in algorithmic implementation
2. Current LLM constraints may restrict full potential

Future Work

1. Develop enhanced optimization modules
2. Investigate collective intelligence among multiple Gödel Agents
3. Implement safety measures for autonomous agents

Additional Notes

- The Gödel Agent framework represents a significant advancement in autonomous AI
- It challenges traditional AI design paradigms by fully exploring autonomous design spaces
- This work sets a trajectory for future research in self-improving AI systems